

The Basics of XML

What is XML?

Extensible Markup Language (XML) is an abbreviated version of Standard Generalized Markup Language (SGML), for the exchange of structured documents over the Internet. Unlike HTML, XML readily enables the definition, transmission, validation, and interpretation of data between differing computing platforms and applications. XML permits people in a specialized field, such as chemistry, finance, or environmental data collection, to develop XML schema that define the markup language for the exchange of specialized data unique to their fields. XML schema is the primary data format supported for data exchange by the State/EPA Environmental Information Exchange Network (Exchange Network).

XML is extensible, meaning a developer can *extend* the language by devising new tags to describe and share data in any specialized way desired as long as the new tags follow the XML syntax defined by the W3C XML specification. XML is very useful for organizations that do not share but need to develop a common data exchange format. Its extensibility provides flexibility in developing exchange formats in XML schema, provided all partners agree on the data format and definitions of the data it contains.

How does XML work?

XML is not a programming language like C++ or Java, but a markup specification language. A browser or other application must read the XML document in order to make it do something.


Development of XML documents begins with the identification and definition of the data elements that will be displayed or exchanged. The data elements are defined using tags that indicate what a data element represents. For example, in the simplest explanation, a person's name might contain three tags:

```
<Name>
  <First>Fred</First>
  <MiddleInitial>M</MiddleInitial>
  <Last>Smith</Last>
</Name>
```

Developers create a schema that contains the tag name, their data formats, and the relationships to one another.

What is an XML schema?

An XML schema is a single file or collection of files that serve as the framework for defining the data content, format and structure of an XML document. A well written schema expresses agreed upon vocabularies and allows computers to validate the data against the rules written to describe the data. It provides a means for defining the structure and content of XML documents. These agreed vocabularies and data structures are designed by the EPA's exchange partners and monitored for consistency with data standards and best practices by the State/EPA Technical Resources Group (TRG).



While a schema describes the content and structure of an XML document, an XML instance document contains the actual data. The schema and instance document are always separate, and the instance document is frequently validated against the schema to ensure the contents and structure are correct. XML schema can be combined and reused, so that one schema can reference another. When creating schema for new data flows, schema developers should look for existing schema that may describe a portion of their data flow. An example may be facility data described in EPA's Facility Registry System (FRS) schema. Exchange Network schema can be found on the Network's XML registry at www.exchangenetwork.net.

How does XML benefit states and EPA programs?

By agreeing on shared vocabularies and data formats, and reusing existing schemas, states and EPA programs can develop shared schemas. This allows different programs at EPA and other federal, state, and local agencies to share and analyze electronic data sets that cross state and program boundaries. Some stakeholders may have a need or desire to analyze data from multiple EPA programs, such as the Air Quality System (AQS) or Toxics Release Inventory (TRI). Duplicate data collection can also be reduced by standardizing data types and formats across programs. In addition, XML can establish additional rules for submitting data that would not be possible with the traditional hard-copy submission process, such as validation controls that increase data quality and reduce the manual correction process on the receiving end.

How does this impact current program data systems?

Instead of EPA's exchange partners (including state, regional and local agencies) translating their data into current EPA program format (e.g., column delimited, flat files, ASCII), they are able to translate their data to the defined XML format based on the schema for the particular data flow, such as RCRAInfo or TRI. With this mapping and schema, most modern databases will create the necessary XML instance document to send the data. Similarly, most modern databases will accept XML data from mapped schema.

No determination has been made to eliminate the EPA program offices' existing formats, but exchange partners will have the option to submit data in XML via a Web page upload process, or a machine-to-machine Web services transfer, as data flows become available.

How can I find out more about XML or the Exchange Network?

World Wide Web Consortium
<http://www.w3.org>

Environmental Information Exchange Network
<http://www.exchangenetwork.net>

EPA Central Data Exchange
<http://www.epa.gov/cdx>

EPA Data Standards
<http://www.epa.gov/edr>

EPA Network Grants Program
<http://www.epa.gov/neengprg/>